Title
Planned Missing Designs and Diagnostic Classification Models

Author
Suh, Yon Soo

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Planned Missing Designs and Diagnostic Classification Models

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Yon Soo Suh

2022

ABSTRACT OF THE THESIS

Planned Missing Designs and Diagnostic Classification Models

by

Yon Soo Suh

Master of Science in Statistics

University of California, Los Angeles, 2022

Professor Yingnian Wu, Chair

Missing responses are often inevitable in assessments, whether they are intended or not. The problem is not with the missing data itself but how it is dealt with. In fact, as in the case of planned missing (PM) data designs, missing data can even be used to our advantage to promote cost-effectiveness and design efficiency in test development. Over the years there has been active research on the impact of, treatment for, and use of different kinds of missing data on psychometric models for assessments with a focus on first classical test theory (CTT) and then item response theory (IRT) models. IRT models have become one of the most popular statistical models for psychometrics and they have been widely used in many educational settings. Nonetheless, in an era of accountability in schools with increased emphasis on providing detailed and formative feedback on individual students, a different flavor of IRT models, coined diagnostic classification models (DCMs), have been fast gaining popularity in the same settings. DCMs specialize in the classification of respondents according to their mastery of a predefined set of underlying cognitive processes called attributes and is well-suited for

obtaining diagnostic information about individual attributes as well as their combinations. However, there is scant research on the impact of any kind of missing data that has been tailored specifically to DCMs. As a step toward filling this gap, this study investigates the effect of using a maximum likelihood (ML)-based approach for treating missing data assumed to be missing completely at random (MCAR) under specifically PM design scenarios using simulated data. Key factors of the type of PM design, the number of attributes, the structural model of DCMs, and sample size were experimentally manipulated to examine the extent to which item parameters of DCMs can be recovered and to compare the effects of various design factors. This project adds to the empirical knowledge base on the statistical properties of DCMs in the face of missing data, which in turn are expected to improve the design and use of DCMs in practical settings.

The thesis of Yon Soo Suh is approved.

Chad J. Hazlett

Minjeong Jeon

Yingnian Wu, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

Introduction

1.1 Background

In an era of accountability in schools, monitoring students' learning progress and providing detailed and formative feedback is an essential issue for students and stakeholders alike. However, obtaining actionable information for diagnostic decision-making or gaining insight into underlying cognitive processes is not easily available from the current popular set of psychometric models of item response theory (IRT; de la Torre & Douglas, 2004; Henson et al., 2009). Conventional IRT models assume that a respondent's latent abilities are continuous and are designed to locate each respondent along that latent continuum quantitatively. They are more suitable for measuring aggregated achievement in broad, summative assessments, and making norm-reference interpretations (Rupp & Templin, 2008).

On the other hand, diagnostic classification models (DCMs) or cognitive diagnostic models (CDMs) were developed specifically to target more fine-grained cognitive processes and provide diagnostic information on their states (Kaya & Leite, 2017; Madison & Bradshaw, 2018). In DCMs, the underlying latent abilities of interest are assumed to be categorical, and the aim is to classify respondents with regard to these categorical latent traits to their most probable mastery profiles. These mastery profiles show which combination of latent abilities respondents have mastered or not mastered and provide feedback on target areas to work on. Accordingly, they have been gaining favor over the recent years, particularly in education, in light of the need for classification-based inferences regarding students' learning, with a myriad of research

on these types of models following suit. Nonetheless, one area that remains unexplored is the impact of missing data on DCMs.

Missing data constitute a source of concern for all kinds of statistical analyses, and the field of psychometrics, where DCMs live, is no exception. If missing data and its mechanisms are not appropriately accounted for, as in traditional methods such as listwise or pairwise deletion and mean substitution, missing data can result in problems such as lack of statistic power, biases in population parameter estimates, and thus less generalizability of findings, inaccurate standard errors of parameter estimates, and distortions in parameter distributions (Enders, 2001b; Graham, 2009; Little et al., 2013; Peng et al., 2006; Schafer, 1997).

Rubin's (1976) missing data mechanism classifies missing data into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing data under MCAR and MAR are considered more or less unproblematic for likelihood-based inferences (Mislevy & Wu, 1996; Rubin, 1976) because the process causing the missing data is known so that the missingness can be controlled for. Thus, the missingness is deemed ignorable (Köhler et al., 2015; Rose et al., 2017). This is thanks to the development of modern missing data analysis tools such as multiple imputation (MI; Rubin, 1987; Schafer, 1997) and maximum likelihood (ML) based approaches such as full information ML (FIML). Using these methods, the aforementioned problems do not occur, and valid inferences can be made under the assumption of missing data of at least MAR. However, when there is a relationship between the underlying process for missing responses and the measured attribute, the missingness is MNAR. Then it is not ignorable and there will be consequences if treated

as such (i.e., biased inferences). How to deal with MNAR is still an ongoing topic of debate without a clear consensus.

In the educational realm and particularly, missing data of all three types is common, particularly in assessment settings. While MNAR, which often occur due to respondents intentionally skipping or omitting items, is still problematic, the use of FIML and MI procedures in conjunction with IRT models have allowed researchers to obtain unbiased parameter estimates and reasonable standard errors when data can be treated as at least MAR. In fact, education is a field that has particularly used ignorable missing data and their modern treatments to its advantage in the form of various planned missing (PM) data designs. PM designs intentionally include data missingness to promote the cost-effectiveness and efficiency of research designs with a focus on shortening the length of measures to reduce respondent fatigue, assessment time, and data collection costs (e.g., Conrad-Hiebner et al., 2015; Graham et al., 2006; Little & Rhemtulla, 2013; Zhang & Yu, 2021). PM designs have been especially instrumental in the case of large-scale assessments, where without the use of PM designs, student response data collection at this massive scale would be nearly impossible (Revelle et al., 2021). Accordingly, there has been a lot of research on missing data in psychometrics, including PM designs. However, as noted above, most of the research has been centered around IRT models (and in the past classical test theory (CTT)). Therefore, there is a severe lack of research on up-and-coming DCMs. Although we can theoretically expect the FIML or MI techniques to function similarly in the case of DCMs, there is currently no research that empirically shows this, let alone investigates the boundaries or conditions of acceptable levels for model design factors, such as the amount of missingness or DCM specific factors like the number of attributes or sample size.

1.2 Research Objectives

This project sought to add to the literature on DCMs by focusing on the effects of missing data. In particular, PM designs were of interest because of their potential for use with DCMs applications that involve an increasingly larger number of attributes. For model identification and stable estimation of DCM model parameters (Fang et al., 2019; Hartz, 2002), researchers such as Hartz (2002) recommends at least three items per attribute (five have also been proported by Jang (2009)), among other requirements. As the number of attributes grows larger, so does the number of needed items. Furthermore, there is also an increased call for diagnostic feedback in even large-scale settings where PM designs are the default. The purpose of this study was twofold. First, the performance of the two common PM designs–the common form design and balanced incomplete blocks (BIB) design–is explored. The results are compared both in terms of the recovery of "true" parameters as well as DCMs with no missing data. Second, in this process, important factors for DCM estimation, specifically, the number of attributes, structural model formulation, and sample size, were considered, especially with regard to their impact on missing data. The specific research questions were:

1) Are there differences in item parameter recovery of DCMs due to missing data arising from PM data designs?

2) Are there differences in item parameter recovery of DCMs using PM data designs following the number of attributes?

3) Are there differences in item parameter recovery of DCMs of PM data designs depending on how the structural model is formed?

4) Are there differences in item parameter recovery of DCMs of PM data designs according to sample size?

CHAPTER II

Literature Review

2.1 Missing Data and Psychometrics

2.1.1 Missing Data Mechanisms

In his seminal work, Rubin (1976) introduces a taxonomy based on the missingness nature (i.e., ignorable versus nonignorable) and identified three different underlying missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Let us consider responses y_{ij} where respondents j ($j = 1, \dots, J$) responds to i ($i = 1, \dots, I$) dichotomous items. A missing data indicator d_{ij} is defined as

$$d_{ij} = \begin{cases} 0, & if\ y_{ij}\ is\ not\ missing\ (observed) \\ 1, & if\ y_{ij}\ is\ missing \end{cases}. \quad (1)$$

The observations and missing data indicators are collected in $J \times I$ matrices $\boldsymbol{Y}$ and $\boldsymbol{D}$, respectively. $\boldsymbol{Y}$ is partitioned into an observed data part $\boldsymbol{Y_{obs}}$ and a missing data part $\boldsymbol{Y_{miss}}$ where

$$\begin{aligned} d_{ij} = 0 &\Rightarrow y_{ij} \in \boldsymbol{Y_{obs}} \\ d_{ij} = 1 &\Rightarrow y_{ij} \in \boldsymbol{Y_{mis}} \end{aligned} \quad (2)$$

The complete data matrix is hence $\boldsymbol{Y} = (\boldsymbol{Y_{obs}}, \boldsymbol{Y_{miss}})$. Let ι denote the parameter vector of the model of $\boldsymbol{Y}$, and φ the parameter vector of the model of $\boldsymbol{D}$ with the respective parameter spaces being Ω_{ι} and Ω_{φ}.

The defining features of the different missing data mechanisms are the unconditional and conditional stochastic dependencies between $\boldsymbol{Y}$ and $\boldsymbol{D}$. The missing-data mechanism is MCAR if

$$\boldsymbol{D} \perp \boldsymbol{Y} \tag{3}$$

In other words, missingness is stochastically independent of the items.

The missing-data mechanism is MAR if

$$\boldsymbol{D} \perp \boldsymbol{Y_{miss}} | \boldsymbol{Y_{obs}} \tag{4}$$

meaning conditional stochastic independence of the missing data part given the observed part or missingness does not depend on $\boldsymbol{Y_{miss}}$. Rubin's ignorability principle (i.e., the missing data mechanism is "ignorable" or "uninformative" for direct likelihood inferences and Bayesian inferences) applies if two conditions hold: a) the missing data are (at least) MAR and 2) the parameters of the data distribution, ι , are distinct from the parameters of the model for the missing data mechanism, φ. If both conditions are satisfied, with the former being more important, the distribution of $\boldsymbol{D}$, is ignorable so that it is possible to get unbiased parameter estimates utilizing only the part of the data that has no missingness (when data are MCAR) or by considering the conditional distribution (when data are MAR).

Lastly, the missing-data mechanism is MNAR if

$$\boldsymbol{D} \not\perp \boldsymbol{Y_{miss}} | \boldsymbol{Y_{obs}} \tag{5}$$

In this case, the missing data are nonignorable or informative (i.e., $\boldsymbol{D}$ provides added information about $\boldsymbol{Y_{miss}}$ over and above $\boldsymbol{Y_{obs}}$) and ignoring them leads to biased parameter estimates. In this case, a joint model of ($\boldsymbol{Y_{obs}}$, $\boldsymbol{D}$) is required for unbiased

estimation of ι. In other words, the missing data mechanism itself has to be modeled with sound reasoning behind the missingness and what the plausible values may be. For example, modeling the process that caused the missing data (Heckman, 1979) via locating and including the explanatory variables of the missing data process and making inferences concurrently on the missing data model and the intended model for the observed data can mitigate such adverse effects (e.g., Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999).

Missing data mechanisms have discrepant consequences for statistical results when they are handled improperly, as in the case of traditional approaches such as pairwise or listwise deletion. The consequences tend to become more severe in the order of MCAR, MAR, and MNAR. Deleting missing data that is MCAR prior to analyses often results in a loss of power due to the loss of sample size. Nonetheless, it does not bias parameter estimates (Enders, 2011; Graham, 2009; Nakagawa & Freckleton, 2010). On the other hand, when missing data are MAR or MNAR, simply excluding the missing data will result in biased parameter estimates along with bringing about a loss of power.

2.1.2 Statistical Procedures for Data Assumed to be At Least Missing At Random (MAR)

Most modern statistical techniques for dealing with missing data rely on the pattern of missingness being ignorable and, thus, either MCAR or MAR. In these cases, they can recapture the data that was missing by design and improve bias and coverage in parameter estimates without losing statistical power (Enders, 2001; Nakagawa, 2017; Schafer & Graham, 2002). Such model techniques can be classified as either MI methods or ML methods such as FIML. Extensive research has shown that both methods can

provide valid conclusions in the face of missing data; although there can be cases where one is favored over the other.

As the name implies, multiple imputation (Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002) involve imputing multiple plausible values for the missing data, as if they had been observed, and creating multiple complete datasets with the purpose of overcoming the problem of biased uncertainty of single imputation methods (Little et al., 2014). MI proceeds using a two-step approach where first a designated set of complete data sets are generated via the imputation of missing value using other observable variables. In the second step, the multiple generated data sets are each analyzed using standard methods as if they were complete case data sets. The resulting parameter estimates and standard errors are then pooled using Rubin's rules (Little & Rubin, 2002; Nakagawa, 2017; Rubin, 1976; Schafer, 1997). Different kinds of imputation techniques are currently available other than MI, which will be touched upon in the context of psychometrical settings.

MI methods are also called "data-based" missing data methods because they achieve results like FIML through the use of multiple datasets. Contrarily, ML approaches are called "model-based" methods as approaches such as FIML allow for the handling of missing data and parameter estimation in a single step as long as ML estimation is possible. Is important to note that the FIML estimator does not impute any missing values but directly estimates model parameters and corresponding standard errors (Enders, 2001). The basic idea is that partially complete observed responses can supplement the loss of information due to missing data (Little et al., 2014). ML methods use all of the available observed data to estimate parameter estimates and standard errors by defining a case-wise likelihood function for each row of complete data, which

are summed together and maximized. This maximizes the probability that the observed data are from the population implied by the parameter estimates (Noble & Nakagawa, 2021).

2.2 Missing Data and Psychometrics

2.2.1 Classification of Missing Data

In the realm of psychometrics, following Rubin's framework, missing data are first classified as planned or unplanned, and then further divided within each category. In total, there are largely four categories of missing data (Lord, 1974: Pimental, 2005). Planned missing (PM) data is also called data that are missing by design as they occur because of intended decision-making and are, therefore, under the control of test administrators or researchers. There are two subtypes within planned missing data. The first category is missing data in *a priori* fixed incomplete tests and calibration designs such as matrix sampling using booklets. Usually, these kinds of PM designs are employed to promote the cost-effectiveness or efficiency of the measurement, which will be explained in more detail later on. As the missing data are *a priori* fixed, it is inherently independent of $\boldsymbol{Y_{miss}}$ or $\boldsymbol{Y_{obs}}$, so that the missing data in MCAR. Thus, ignorability trivially holds. The second subtype of PM data is missing occurring because of instrument characteristics. Examples include data from response-contingent designs like two-stage or multistage testing and computerized adaptive testing (CAT). Because the items administered (or conversely, not administered) are entirely determined by the observed responses, they are independent of any unobserved responses. Therefore, the missing data mechanism is MAR. Like this, these two types of PM data designs produce ignorable missing data.

Unplanned missing data also has two subtypes as well. Although both subtypes result from responses like “don’t know” or “not applicable,” they are.divided by whether the missing data can or cannot be ignored depending on the cause of the missingness. If the missing responses do not depend on the latent variable being measured, these responses can be considered MAR and, therefore, ignored in the analysis. On the other hand, if the “don’t know” or “not applicable” responses because there is a relationship between the propensity of a response to be missing and the response, they are resulting from a MNAR mechanism. Examples of MNAR responses include low-ability respondents failing to produce a response and skipping items (i.e., item omission) or items not being reached due to time constraints. In these contexts, the probability of omitting and/or not reaching an item depends not only on the item characteristics and latent trait(s) of interest but additionally on unobserved, latent variables like the missing distribution. As such, the assumption of MAR is violated and treating these responses as ignorable missing can produce the problems mentioned above (Holman & Glas, 2005; Rose et al., 2010).

2.2.2 Methods of Handling Missing Data

Existing methods for dealing with missing responses in testing or survey settings consist of classical, imputation-based, and model-based approaches under MNAR (Finch, 2008). Classical approaches have four subtypes. The first subtype is to ignore missing responses, that is, treat as though not given, and assume at least MAR so that modern missing data techniques such as FIML can be applied. Or missing values can be scored as incorrect, which involves making a deterministic decision to ignore an examinee’s positive probability to solve the item, conditional on ability. Yet another deterministic imputation method is to score missing responses as fractionally correct and give scores

such as the reciprocal of the total number of item categories (i.e., 1/(number of item response options). Lastly, a mix of the subtypes above can be utilized as in two-stage procedures where missing responses are ignored when calibrating item parameters but then are considered incorrect when scoring persons. Variants of these classical approaches are the default in large-scale educational assessments such as NAEP, NEPS, PISA, and TIMSS (Köhler et al., 2015).

Item nonresponse imputation methods consist of three subtypes: deductive imputation, deterministic imputation, and stochastic imputation. In deductive imputation, missing data are imputed using other known information. In deterministic imputation, a predicted or specific value is used. In stochastic imputation, randomness is incorporated in the imputing process. Deductive imputation is seldom feasible in assessment or survey settings, but different flavors of both deterministic and stochastic imputations are possible. Deterministic imputation includes methods such as unconditional mean imputation, person mean imputation, and regression imputation. Classical approaches can be considered deterministic imputation as well. Stochastic imputation incorporates uncertainty in the imputation process through a randomness variable. It involves methods such as stochastic regression imputation and MI fall in this last category. Stochastic imputation methods are considered superior to other imputation methods (Finch, 2008).

Model-based approaches, which consist of latent and manifest approaches (Rose et al., 2010), have been proposed to handle MNAR data. Mainly three methods fall under latent variable approaches. One is treating missing value code as a separate response category in addition to responses for observed items in a nominal response model. While this approach avoids adding another dimension to the model, in this case, the

assumption is that responses and nonresponses are all related to the same dimension(s) (Moustaki & Knott, 2000). The other two approaches both attempt to identify and model the missing-data mechanism by employing a measurement model for a latent response/missing propensity that is jointly estimated with a measurement model using the observed data for the ability of interest. For this purpose, the former of the two approaches computes response propensities based on the idea of employing propensity score methods (i.e., fitting logistic or probit regression models with binary item response/nonresponse variable regressed on a set of covariates) to weight item responses and respondents accordingly and obtain adjusted estimates (O'Muircheartaigh & Moustaki, 1999). The latter approach is the most widely applied method of a latent variable or IRT model with (at least) two latent dimensions, one for the latent response/missing propensity and the other for ability to be essentially a multidimensional IRT (MIRT) model where the correlation between the latent traits are estimated (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999).

2.3 Planned Missing (PM) Data Designs

PM data designs involve intentional and strategic use of missingness in the data to promote the efficiency of a study or assessment (Graham et al., 2006; Little & Rhemtulla, 2013). Examples include subjects who don't provide responses to certain items or measures or at some time points. In all PM designs, random assignment of missingness is essential because then the missing data are by definition MCAR. Under MCAR, there is no bias in estimated parameters, and the diminished power can be recovered by modern treatments for missing data (e.g., FIML and MI; Little et al., 2014). In short, PM designs capitalize on the fact that missing data of MCAR or MAR process

(especially the former) can be easily recaptured via modern techniques (Enders, 2010; Graham et al., 2006; Little & Rubin, 2002).

In PM data designs, researchers and developers are in control of missingness, granting them control over the missing data mechanism and are further using it to their advantage. Strategic implementation of PM designs allows researchers and developers to optimize various costs under constraints to produce an efficient and cost-effective study or assessment design with the best possible outcome. The benefits of PM designs include having to collect less data on a given participant. For instance, multiform designs (Graham et al., 1994), which are the focus of this study, result in an increase in the number of measures or items given to each participant without increasing their burden. This results in the shortening of lengths of instruments and measures without compromising validity and can also reduce respondent fatigue to promote validity (Graham et al., 2006; Noble & Nakagawa, 2021). In the case of repeated measurements, PM designs can help achieve the number of measurement occasions needed with fewer repeated measures from an individual (Hogue et al., 2013; Wu et al., 2016) to reap similar benefits.

Furthermore, PM designs can be used to combine measurements of different variables to counteract the "fallacy of the factorial design" problem (Betini et al., 2017). It is also possible to combine measurements of the same variable using different methods to boost statistical efficiency (Hogue et al., 2013). For example, under a PM design, only a subset of participants can be given an extensive gold-standard measure while the rest are offered a cheaper alternative. This can offset the cost of the study while increasing sample size and, thus, statistical power (Graham et al., 2006; Little et

al., 2017). Like this, PM designs can improve statistical power along with model convergence relative to complete data cases (Little & Rhemtulla, 2013).

Furthermore, PM designs can aid in counteracting problems of MNAR (Little & Rubin, 2002; Nakagawa, 2017) through the reduction of circumstances where missingness can occur unintentionally or the addition of variables that can increase the possibility of correlation with missingness so that the missing data becomes MAR. This will help to correct for nonignorable missingness which are issues for complete case analyses.

2.2.1 Two Types of Planned Missing Data Designs

A multitude of different PM designs is possible across different study scenarios. Some popular types of PM data designs include the multiform questionnaire protocol, the two-method measurement model, and the wave-missing longitudinal design (Noble & Nakagawa, 2021). As noted above, in this study, the focus is on multiform designs. Multiform designs are also referred to as split-questionnaire or matrix sampling designs. In this study, I will, from now on, use the term matrix sampling following the convention in psychometrics (earlier literature uses the term item-sampling). Matrix sampling (Shoemaker, 1973) is a sampling design that samples both examinees and items. In other words, samples of items are given to samples of examinees. Various types of matrix sampling designs exist (Graham et al., 2006) that try to improve the simplest matrix sampling by giving different items to different respondents (see Shoemaker (1973) for more detail). Most matrix sampling designs divide an assessment or measure subsets or "blocks" that ensure coverage of the subscales or constructs and administer one or more blocks to one or more random subsamples of the respondent pool. To reiterate, a block denotes a set of items or variables given to respondents. A set of blocks arranged in a

specific manner refer to an assessment or test "form." That is, a test form is the actual set of items given to participants (Gonzalez & Rutkowski, 2010). It is also often called a "booklet." Two specific types of matrix sampling explored in this thesis were the two-form design (Adams et al., 2013) and a balanced incomplete block (BIB) spiral design.

Two-form design (Adams et al., 2013) is a kind of three-form design developed with the objective of having more items or variables than that could be answered by an individual respondent. Also, the design needs to be able to estimate all correlations as well as means and variances of those items or variables. The basic design consists of a primary block, often denoted X, that contains a set of common items that are assigned to all participants. The other items or measures are also subsetted into blocks which are given to only a random subset of respondents. Like this, the various blocks are utilized in building shorter forms via concatenation of the common block (X) and the blocks of item subsets (Little et al., 2017). Excluding the common block, the two-form design has two rotated blocks of item subsets. In total, three mutually exclusive blocks of items are created. An example of the two-block design is given in Table 2. 1.

Table 2. 1 Two-Form Design with Common Block X

Form	Item Block		
	X	A	B
1	1	1	0
2	1	0	1

Note. 1 = item block assigned; 0 = item block not-assigned.

Balanced incomplete blocks (BIB) spiral designs also involve blocks of items or variables, and through this design, the means of all items and variables can be estimated as well as correlations among all pairs of items and variables (Gonzalez, Rutkowski, 2010; Graham et al., 2006). First suggested by Lord (1965) for use in the context of multiple-

matrix sampling, many variants of this design (e.g., partially BIB (PBIB) designs) are in use today. The design is "balanced" in that every item block appears an equal number of times in all block positions. Accordingly, the number of respondents for each item block and each pair of item blocks is equal. An example of the BIB with three forms and three item blocks is given in Table 2. 2. As can be seen by summing across rows for forms, each item block appears twice. It is also possible to see from Table 2. 2 that each block is paired once with every other block (i.e., A with B, A with C, B with C).

Table 2. 2 Balanced Incomplete Block Design

Form	Item Block		
	A	B	C
1	1	1	0
2	1	0	1
3	0	1	1

Note. 1 = item block assigned; 0 = item block not-assigned.

2.3. Diagnostic Classification Models (DCMs)

DCMs are intent on estimating an examinee's latent abilities, coined *attributes*, on a discrete scale in terms of varying statuses of mastery or attainment (e.g., mastery versus non-mastery) of each attribute. The main purpose of DCMs is to "diagnose" or assign the most likely attribute mastery patterns (i.e., *attribute profiles*) for each student, that is, the combination of attributes they have mastered or not mastered. Thus, DCMs make it possible to measure specific knowledge structures and processing skills (i.e., attributes) and provide multiple criterion-referenced interpretations and diagnostic feedback about the mastery and non-mastery of attributes at even very fine-grain size levels (Leighton & Gierl, 2007; Rupp & Templin, 2008).

2.3.1 Characteristics of Diagnostic Classification Models (DCMs)

DCMs are probabilistic, confirmatory multidimensional latent-variable models where the latent variables are considered to be discrete. Most of these characteristics, excluding the key distinction with regard to the distribution of the latent traits, can be found in other IRT or factor analytic (FA) models as well.

As latent variable models, DCMs make the essential distinction between manifest or observed variables and latent or unobservable models that are driving the responses to the observed variables. Latent variables are thus the true parameters of interest. In DCMs (as well as IRT models) the observed variables or items are assumed to be categorical (e.g., dichotomously and polytomously scored item responses) and are related to the underlying categorical latent variables via a probabilistic model. Elaborating, DCMs assume a probabilistic model for the observed categorical responses according to 1) a vector of categorical latent factors called attributes and 2) an item-specific mapping between the attributes and the probabilities of observing a certain response category (e.g., correct or positive response for dichotomously scored items; Duck-Mayr et al., 2020). Based on this probabilistic model, individualized discrete attribute mastery profiles are generated as a series of probabilities on mastery classifications (Rupp et al., 2010) from which the most probable profile is selected.

Regarding the number of latent traits, DCMs generally involve multiple latent dimensions to be more in line with MIRT models as opposed to unidimensional IRT models with only one latent trait (Rupp & Templin, 2008). In fact, DCMs can be considered as a special type of MIRT model modified to handle latent abilities that are categorical in nature as opposed to continuous. In DCMs, the continuous latent traits of IRT models are reconceptualized as a set of attributes following categorical distributions

denoting mastery levels (Cai et al., 2016; Rupp & Templin, 2008). Usually, multivariate Bernoulli distribution (MVB) where 1 denotes mastery (presence) of an attribute and 0 means non-mastery (absence) is assumed. Moreover, DCMs are typically concerned with finer-grained subdomains as opposed to more broad and general latent abilities targeted in IRT models. That is, DCMs produce multiple qualitative classifications on several finer-grained dimensions, while IRT provides quantitative scaled scores on broad domains (Bradshaw, 2017; Rupp et al., 2010).

DCMs are also confirmatory in nature because the attributes and how they interact both with each other and in relation to items should be specified *a priori* based on solid theory. The incidence matrix, called a Q-matrix (Tatsuoka, 1983), maps items to attributes where rows represent items and columns represent attributes. Elements of the Q-matrix (refer to Table 3. 2 for an example) are assigned 1 if an item is measured by an attribute and otherwise 0, making it similar to loading structures in confirmatory FA and IRT models (Rupp & Templin, 2008). Nonetheless, these models typically assume simple loading structures where each item only loads on very few dimensions (Rupp & Templin, 2008). Contrarily, as DCMs typically deal with fine-grain constructs, there is an increased likelihood for complex association patterns (Cai et al, 2016), and thus complex structure Q-matrices where items depend on several attributes.

2.3.2 Formulation of Diagnostic Classification Models (DCMs)

DCMs are essentially latent class analysis (LCA) models with some restrictions. In LCAs, observable categorical response variables are connected to discrete latent traits with the purpose of finding the underlying latent class profiles (Rupp & Templin, 2008). Corresponding to this, in DCMs, item responses are used to measure the mastery or non-

mastery of a set of attributes and to "diagnosis" the most likely attribute mastery patterns (i.e., attribute profiles) for each respondent.

Consider N respondents $j = 1, \cdots j$, each of whom responds to I dichotomously scored items of an assessment ($i = 1, \cdots I$) with A dichotomous attributes. Let $\boldsymbol{y}' = (y_1, y_2, \cdots, y_J)$ denote the vector of I variables. Responses to the items belong to an I-way contingency table with a total of $R = 2^I$ cells that denote the possible response vectors $\boldsymbol{y'_r} = (c_1, c_2, \cdots, c_J)$ where $r = 1, \cdots R$ and $c_j = \{0,1\}$. $c_1, c_2, \ldots, c_I$ represent the response categories of $y_1, y_2, \ldots, y_I$, respectively.

Under the general LCA framework, the probability for an item response vector $\boldsymbol{y}'_r$ can be specified as:

$$P(\boldsymbol{Y}_r = y_r) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} P(Y_{ri} = y_{ri}|c_r) = \sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{ic}{}^{y_{ri}}(1 - \pi_{ic})^{1-y_{ri}} \quad (6)$$

v_c is a mixing probability $\left(\sum_{c=1}^{C} v_c = 1.0\right)$ and denotes the probability that a randomly selected individual belongs to the latent class or profile c. π_{ic} is the probability of a correct response to item i given membership in attribute profile c. y_{ri} is respondent r's dichotomously scored response to item i. The product across items is based on the conditional independence assumption of DCMs (and IRT models), where within a latent class, responses to items are considered to be independent. Like this, the unconditional or marginal probability of a particular item response vector $\boldsymbol{y}'_r$ is a weighted sum over all conditional item response probabilities with class probabilities as weights. The number of attributes, attribute-attribute relationships, and item-attribute relationships are predefined confirmatory restrictions that the DCM imposes on the general LCAs,

making DCMs restricted LCAs via the Q-matrix and specific parameterizations of the v_c and π_{ic} parameters (Rupp et al., 2008).

Measurement Model

The portion of Equation (6) with the product term (i.e., $\prod_{i=1}^{I} \pi_{ic}{}^{y_{ri}}(1-\pi_{ic})^{1-y_{ri}}$) is called the measurement model. It specifies how observed item responses are related to attributes and is therefore concerned with the estimation of π_{ic} or the probability of a correct response to an item for a respondent r in latent class c. A plethora of different DCM variants has been developed with a focus on the measurement model via condensation rules, which dictate how multiple attributes are "condensed" to produce a response to an item response (Rupp & Templin, 2008). Simple structure items where only one attribute is measured by an item are not affected by such rules to give the same results regardless. Conjunctive, disjunctive (Rupp, & Templin, 2008), and additive condensation rules (de la Torre, 2011) are commonly employed rules (Ravand & Baghaei; 2020) contingent on the assumptions of the compensatory versus non-compensatory relationship among the attributes in question (Rupp et al., 2010). Compensatory DCMs propose that the lack of one or more attributes can at least be partially or completely offset through the possession of other required attributes, while non-compensatory models assume that a deficit in one attribute cannot be compensated by the presence of other required attributes.

The deterministic-input, noisy-and-gate (DINA) model is the most popular fully non-compensatory DCM, which follows a conjunctive condensation rule postulating that all relevant attributes must occur in conjunction with each other. The deterministic input, noisy-or-gate (DINO) model is the most well-known disjunctive model (Ravand & Baghaei, 2020) and assumes that the maximal probability of a positive or correct item response

can be achieved through mastery of at least one of the measured attributes (Henson et al., 2009). Thus, it can be considered an extreme compensatory DCM as a single attribute can fully offset the non-mastery of all others (Rupp et al., 2010). The compensatory reparameterized unified model (C-RUM) is a widely used additive model (Rupp et al., 2010) that supposes mastery of each related attribute to an item leads to an increase in the probability of item endorsement independent of the mastery statuses of others. The nature of the C-RUM model allows mastery of attributes to at least partially compensate for non-mastery in others.

The main distinction among different measurement models for DCMs is based on how the latent attributes are combined to generate a correct response to an item. Many of these models can be subsumed and brought under a unified framework using general DCMs such as the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009). This study also applies this framework due to the flexibility it affords us in model building, parameter estimation, and model evaluation. As can be inferred by the name, LCDM uses a log-linear framework to parametrize the relationship between examinee attribute mastery and probabilities of correct item responses. Elaborating, the LCDM item response function for π_{ic} is

$$P\left(Y_{ji}=1\middle|\boldsymbol{\alpha}_c\right)=\frac{\exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_c,\boldsymbol{q}_i)\right)}{1+\exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_c,\boldsymbol{q}_i)\right)} \tag{7}$$

$P\left(Y_{ji}=1\middle|\boldsymbol{\alpha}_c\right)$ refers to the probability that respondent j with the attribute profile $\boldsymbol{\alpha}_c$ correctly or positively responds to item i. Item parameters include an intercept of $\lambda_{i,0}$, which is the baseline log-odds of success for individuals with no mastery. The multiplicative portion $\boldsymbol{\lambda}_i^T\boldsymbol{h}(\boldsymbol{\alpha}_c,\boldsymbol{q}_i)$ consists of $\boldsymbol{\lambda}_i^T$ and $\boldsymbol{h}(\boldsymbol{\alpha}_c,\boldsymbol{q}_i)$. $\boldsymbol{\lambda}_i^T$ is a column vector of

simple main effects for specific attributes as well as interaction effects between the attributes for an item i. These parameters give the necessary change in log-odds for item endorsement as attributes and their combinations are added to the model. $\boldsymbol{h}(\boldsymbol{\alpha}_c, \boldsymbol{q}_i)$ is a vector of size $2^A - 1$ that take on values of 0 or 1 depending on attribute mastery indicators $\boldsymbol{\alpha}_c$ and Q-matrix entries $\boldsymbol{q}_i$.

Parametric forms for specific DCMs of conjunctive, disjunctive, as well as additive DCMs, can all be derived under the LCDM framework by constraining the parameters in Equation 7 accordingly (Hansen & Cai, 2013; Henson et al., 2009).

Structural Model

The portion of Equation (6) with the additive term $\sum_{c=1}^{C} v_c$ refers to the structural model of a DCM (as opposed to the measurement part related to item response probabilities). The structural parameters (v_c's) of this part of the DCM refers to the base rate probability of each latent class or attribute profile in the population. They make up the latent attribute space and allow us to examine the marginal distributions of the mastery rate of attributes as well as the correlations among attributes. In the case of A binary latent attributes, the maximum number of v_cs is equal to the number of attribute profiles or 2^A. Estimating all of these possible attribute profiles is denoted as an "unstructured" structural model where $2^A - 1$ parameters (the minus 1 is necessary because of the constraint that the probabilities must sum to unity: $\sum_{c=1}^{2^A} v_c = 1$) are directly estimated.

Although the most flexible, the number of parameters to be estimated increases exponentially with the addition of attributes. Thus, methods for reducing the number of structural parameters are desirable, especially as the number of attributes grows larger.

Different structural models have been suggested in the literature (Rupp et al., 2010). One such structural model is the independence model. As the name suggests, it assumes that attributes are statistically independent. Accordingly, it is one of the most parsimonious models as only A structural parameters are estimated. These structural parameters refer to the population proportions for each attribute.

Nonetheless, attributes are rarely independent, and structural models for the joint distribution of latent attributes have been developed for the dual purpose of reducing model complexity and incorporating hypothesized structures. For this purpose, log-linear structural models (Henson & Templin, 2005) are frequently used due to their flexibility and ability to provide both a "top-down" approach to deciding on the number of structural parameters to be estimated as well as an a priori specified structure (Thompson, 2018). Under a log-linear parameterization, the v_cs are further modeled using a log-linear model containing main effects and interaction effects, not unlike the one used for the LCDM parameterization of the measurement model. The kernel for latent class c is

$$\sum_{a=1}^{A} \gamma_{1,(a)} \alpha_{ca} + \sum_{a=1}^{A-1} \sum_{a'=a+1}^{A} \gamma_{2,(a,a')} \alpha_{ca} \alpha_{ca'} + \cdots + \prod_{a=1}^{A} \gamma_{A,(a,a',\ldots)} \alpha_{ca} \tag{8}$$

$\gamma_{1,(a)}$ are the main effect parameters for each attribute mastered by respondents in a class and the rest are interaction parameters between the mastered attributes (with $\gamma_{2,(a,a')}$ referring to two-way interactions all the way up to the A-way interaction parameter of $\gamma_{A,(a,a',\ldots)}$. A saturated log-linear structural model including parameters up to the A-way interaction term results in the unstructured structural model. Usually, only the main effects and lower-order interaction terms (up to two-way interaction terms) are

modeled for computational efficiency. The estimation of only the main effects is equivalent to the independence model. The addition of the two-way interaction terms allows for correlations between estimates to be estimated (Thompson, 2018)

Two models involving tetrachoric correlations: the unstructured tetrachoric model (Hartz, 2002) and the structured tetrachoric model (de la Torre & Douglas, 2004) have also been suggested where discretized multivariate normal distributions are imposed for the attributes. In the former parameterization, the full tetrachoric correlation matrix for all attribute pairs is estimated without additional constraints on the correlation patterns, making it akin to the log-linear parametrization with only main effects and two-way interactions. The latter tetrachoric parameterization places additional constraints on the tetrachoric correlation matrix to make it more structured and to further reduce the number of parameters. Such structured tetrachoric parameterizations can include a higher-order factor model where mastery of the attributes is considered a function of one or more higher-order continuous latent variables.

Imposing a higher-order structure on υ_c where the higher-order, continuous latent traits θs are regressed on the attributes is also popular. The probability of mastering each attribute then depends on a respondent's location on this higher-order dimension. If we assume that the mastery of a set of skills for a respondent is related to a unidimensional trait θ, and assume conditional independence of the attributes given θ, the probability model of $\boldsymbol{\alpha}_c$ conditional on θ is

$$P(\boldsymbol{\alpha}_c \mid \theta) = \prod_{k=1}^{A} P(\alpha_k \mid \theta) \tag{9}$$

where A is the total number of attributes. If the attributes are binary variables, they can be treated as if they were dichotomously scored items and technically, any IRT model may be used for $P(\alpha_k = 1 \mid \theta)$ (Henson, 2013). For example, if a two-parameter logistic (2PL) model is imposed for all attributes, then

$$P(\alpha_k = 1 \mid \theta) = \frac{1}{1 + \exp\left(-(c_k + a_k\theta)\right)} \tag{10}$$

where c_k and a_k are the intercept and slope parameters, respectively, that resemble item easiness and discrimination parameters in traditional IRT models. However, it is important to distinguish that these are higher-order structural parameters and that the higher-order model is being fit to attribute profile probabilities and not the observed item response patterns.

Another possible structure for the attributes that is gaining traction is using attribute hierarchies to specify dependencies among attributes. The idea of attribute hierarchies is far from new, having been suggested from the advent of the rule space model (RSM; Taksuoka, 1983) and attribute hierarchy method (AHM). However, their use has grown substantially since their integration into the LCDM framework (Templin & Bradshaw, 2014) and the rise of learning progressions. Four types of attribute hierarchies prevalent are linear, convergent, divergent, and unstructured hierarchies.

CHAPTER III

Methods

The study examined the effects of PM designs on the item parameter recovery of DCM models using a Monte Carlo simulation study. More specifically, two different types of PM designs of a common item design and a BIB design were compared along with the impact of the number of attributes, structural model formulation, and sample size.

3.1 Data Generation

3.1.1 Fixed Conditions

The fixed conditions for this simulation study are summarized in Table 3. 1 and explained in more detail below.

Q-matrix Design: Not only can Q-matrix design misspecifications have dire consequences, but the number of attributes measured per item or item complexity can substantially affect parameter estimation, classification, and reliability (Lai et al., 2012; Madison & Bradshaw, 2015; Rupp & Templin, 2008). In order to avoid confounding results due to a complex Q-matrix structure where a single item measures multiple attributes, a Q-matrix design of a simple structure, meaning that each item measured exactly one attribute, was considered. The Q-matrix design is given in Table 3. 2.

Number of Items per Attribute: While the number of attributes was varied, the number of items with each attribute was kept equal. Literature on model identification and estimation stability of DCMs, recommends a minimum of three items per attribute as well as having at least one simple structure item per attribute, which was followed in this study.

Measurement Model: The LCDM model described in section 2.3.2 was used due to its flexibility as a general DCM model. As a simple structure Q-matrix was assumed, the form of the LCDM for each model had two types of parameters: an intercept parameter and a main effect parameter for the one attribute which an item was mapped to.

Item Parameters: In the context of DCMs, item quality is measured by item discrimination which can be defined as the difference in item response probabilities for different groups of examinees (Bradshaw & Madison, 2016). In line with this, item parameters were generated so that non-masters of an attribute had probabilities between 0.15 and 0.30 for correctly responding to each item related to that attribute, while masters of an attribute had probabilities between 0.60 and 0.90 for a correct response to the item. The randomly generated item parameters based on these criteria are given in Table 3. 3.

Attribute Base-rates and Correlations: The attribute base rate for each attribute, also called marginal attribute difficulty because it refers to the proportion of examinees who are masters of an attribute, was set equal for all attributes at 0.5. In addition, the tetrachoric correlations reflecting the relationship between factors were also set to a common value of 0.7 (Kunina-Habenicht, Rupp, & Wilhelm, 2012).

Table 3. 1 Fixed Simulation Conditions

DCM measurement model	LCDM
Q-matrix Design	Simple Structure of one attribute per item
Number of Items per Attributes	3
Item Parameters	Non-masters: 0.15 ~ 0.30 probability of correct response Masters: 0.60 ~ 0.90 probability of correct response
Attribute Pre-test Base-rates and Correlations	Attribute Base-rates: 0.5 Attribute Correlation: 0.7

Table 3. 2 Q-Matrix Design

Item	Att 1	Att 2	Att 3	Att 4	Att 5	Att 6	Att 7	Att 8
1	1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0
8	0	0	1	0	0	0	0	0
9	0	0	1	0	0	0	0	0
10	0	0	0	1	0	0	0	0
11	0	0	0	1	0	0	0	0
12	0	0	0	1	0	0	0	0
13	0	0	0	0	1	0	0	0
14	0	0	0	0	1	0	0	0
15	0	0	0	0	1	0	0	0
16	0	0	0	0	0	1	0	0
17	0	0	0	0	0	1	0	0
18	0	0	0	0	0	1	0	0
19	0	0	0	0	0	0	1	0
20	0	0	0	0	0	0	1	0
21	0	0	0	0	0	0	1	0
22	0	0	0	0	0	0	0	1
23	0	0	0	0	0	0	0	1
24	0	0	0	0	0	0	0	1

Note. The box inside Table 3.2 refers to the Q-matrix for the 4 attribute case.

Table 3. 3 “True” Item Parameters

Item		λ_0	λ_1	Item		λ_0	λ_1
1	A1_1	-1.373	3.214	13	A5_1	-0.961	2.048
2	A1_2	-1.246	2.397	14	A5_2	-1.432	3.280
3	A1_3	-1.569	2.866	15	A5_3	-1.098	2.607
4	A2_1	-1.364	3.234	16	A6_1	-1.568	2.559
5	A2_2	-0.874	1.776	17	A6_2	-0.860	1.730
6	A2_3	-1.587	2.488	18	A6_3	-1.421	3.194

7	A3_1	-1.722	3.661	19	A7_1	-1.605	3.062
8	A3_2	-1.553	2.564	20	A7_2	-1.554	3.595
9	A3_3	-0.987	2.414	21	A7_3	-0.888	2.083
10	A4_1	-0.943	2.268	22	A8_1	-0.999	2.532
11	A4_2	-1.225	2.546	23	A8_2	-0.865	1.873
12	A4_3	-1.130	2.883	24	A8_3	-1.372	3.272

3.1.2 Manipulated Conditions

The manipulated factors for this simulation study are summarized in Table 3. 4 and explained in more detail below.

Planned missing Design: The two types of PM designs –a common form design and a BIB design–were considered. The common form design was more specifically a two-form design consisting of a block of common items and two sets of non-common items. The latter BIB design consisted of a total of three forms or "booklets" and three item blocks, with each block appearing twice in each of the two possible positions and each block being paired once with every other block. Despite the difference in the number of forms, the percentage of missingness for each student was kept to the equal reasonable amounts found in the literature of 33% of the complete data, making them comparable. The specific PM designs for this simulation study are in Table 3. 5 for the two-form design and Table 3. 6 for the BIB design. The particular items allocated to the forms are organized in Table 3. 7 for both designs. It is important to note that these designs are only one possibility for item allocation using a specific PM design.

Number of Attributes: Research shows that the most common number of attributes is four. Four and double its number of eight were chosen to represent the average and a large number of attributes.

Structural Model: Various structural models can be imposed on the measurement model. The choice of which model can greatly influence the number of structural parameters to be estimated as the number of attributes grows. For example, for eight attributes, a fully saturated model requires the estimation of $2^8 - 1 = 256$ structural parameters. When a log-linear model with main and two-way interaction effects is imposed, this reduces to $8 + \frac{8 \times 7}{2} = 36$ structural parameters being estimated. If a higher-order one-parameter (1PL) model is used, this further reduces to only $2 + 1 = 9$ structural parameters being estimated. These three structural models–the fully saturated model, log-linear model of lower-order effects, and the higher-order 2PL DCM–are estimated.

Sample Size: As with many statistical models, estimation and, thus parameter recovery is shown to improve following increases in sample size. However, realistically, it is often not possible to have a large sample size. Sample sizes were selected based on the literature on DCMs, where 600 is considered a small sample size, 1200 can be considered the average sample size for DCM studies, and 48000 examinees can be said to be a large sample size.

Table 3. 4 Manipulated Simulation Conditions

		Attribute 4			Attribute 8		
		N=600	N=1200	N=4800	N=600	N=1200	N=4800
Saturated Structural Model	Two-Form Design	S_2F_A4_600	S_2F_A4_1200	S_2F_A4_4800	S_2F_A8_600	S_2F_A8_1200	S_2F_A8_4800
	BIB Design	S_BIB_A4_600	S_BIB_A4_1200	S_BIB_A4_4800	S_BIB_A8_600	S_BIB_A8_1200	S_BIB_A8_4800
	None	S_NA_A4_600	S_NA_A4_1200	S_NA_A4_4800	S_NA_A8_600	S_NA_A8_1200	S_NA_A8_4800
Log-Linear Structural Model	Two-Form Design	LL_2F_A4_600	LL_2F_A4_1200	LL_2F_A4_4800	LL_2F_A8_600	LL_2F_A8_1200	LL_2F_A8_4800
	BIB Design	LL_BIB_A4_600	LL_BIB_A4_1200	LL_BIB_A4_4800	LL_BIB_A8_600	LL_BIB_A8_1200	LL_BIB_A8_4800
	None	LL_NA_A4_600	LL_NA_A4_1200	LL_NA_A4_4800	LL_NA_A8_600	LL_NA_A8_1200	LL_NA_A8_4800
Higher-Order Structural Model	Two-Form Design	HO_2F_A4_600	HO_2F_A4_1200	HO_2F_A4_4800	HO_2F_A8_600	HO_2F_A8_1200	HO_2F_A8_4800
	BIB Design	HO_BIB_A4_600	HO_BIB_A4_1200	HO_BIB_A4_4800	HO_BIB_A8_600	HO_BIB_A8_1200	HO_BIB_A8_4800
	None	HO_NA_A4_600	HO_NA_A4_1200	HO_NA_A4_4800	HO_NA_A8_600	HO_NA_A8_1200	HO_NA_A8_4800

Table 3. 5 Two-Form Design in Simulation Study

Common	Form 1	Form 1
A1_1	A1_2	A1_3
A2_1	A2_2	A2_3
A3_1	A3_2	A3_3
A4_1	A4_2	A4_3
A5_1	A5_2	A5_3
A6_1	A6_2	A6_3
A7_1	A7_2	A7_3
A8_1	A8_2	A8_3

Note. A= Attribute.

Table 3. 6 Balanced Incomplete Block Design in Simulation Study

Form	Item Block					
	A1_1, A2_1, A3_1, A4_1	A1_2, A2_2, A3_2, A4_2	A1_3, A2_3, A3_3, A4_3	A5_1, A6_1, A7_1, A8_1	A5_2, A6_2, A7_2, A8_2	A5_3, A6_3, A7_3, A8_3
1	1	1	0	1	1	0
2		1	1		1	1
3	1		1	1		1

Note. A= Attribute.

Table 3. 7 Item Allocation by Form by PM Data Design

2-Form Design		BIB Design		
Form 1	Form 2	Form 1	Form 2	Form 3
A1_1	A1_1	A1_1	A1_2	A1_3
A2_1	A2_1	A2_1	A2_2	A2_3
A3_1	A3_1	A3_1	A3_2	A3_3
A4_1	A4_1	A4_1	A4_2	A4_3
A5_1	A5_1	A5_1	A5_2	A5_3
A6_1	A6_1	A6_1	A6_2	A6_3
A7_1	A7_1	A7_1	A7_2	A7_3
A8_1	A8_1	A8_1	A8_2	A8_3
A1_2	A1_3	A1_2	A1_3	A1_1
A2_2	A2_3	A2_2	A2_3	A2_1
A3_2	A3_3	A3_2	A3_3	A3_1
A4_2	A4_3	A4_2	A4_3	A4_1
A5_2	A5_3	A5_2	A5_3	A5_1
A6_2	A6_3	A6_2	A6_3	A6_1
A7_2	A7_3	A7_2	A7_3	A7_1
A8_2	A8_3	A8_2	A8_3	A8_1

Note. A= Attribute.

3.2 Analysis Steps

In total, two different types of PM designs, two magnitudes of the number of attributes, three different kinds of structural models, and three varying levels of sample size were of interest. Furthermore, a reference condition built under the same fixed simulation configurations without any possibility of missing data was also considered to serve as the yardstick for comparing results from conditions where missing data was introduced. Including this complete data condition, the total number of simulation conditions was $3 \times 2 \times 3 \times 3 = 54$ conditions.

The analysis proceeded in the following steps: first, complete item response data was generated according to the above simulation conditions. In simulating the PM or matrix sampling designs, item responses from the original complete data sets were deleted for randomly sampled subsets of respondents according to item-by-form allocation guidelines in Table 3. 6. Second, all complete and missing-by-design datasets were calibrated using the LCDM model with a simple structure Q-matrix design. In the case of datasets with missing data, concurrent calibration was used where all items were estimated simultaneously with missing responses treated as is based on FIML estimation. Finally, item parameter estimation was examined by calculating the means of the evaluation criteria over 25 replications of each simulation condition. Thus, the total number of simulations was $54 \times 25 = 1350$.

3.3 Evaluation Criteria

Three types of evaluation criteria were used in this study: model convergence, parameter recovery, and model fit indices (Dai et al., 2021). First, as an initial check, the

model convergence rate across replications was recorded and compared across conditions.

Second, the accuracy of item parameter recovery was evaluated using average bias and average root mean square error (RMSE), which compared the estimated item parameters with their true counterparts averaged across the item parameters as well as over the number of replications. More specifically, each criterion for each parameter was calculated as follows

$$Bias(\hat{\lambda}) = \frac{\sum_{j=1}^{J}(\hat{\lambda}_j - \lambda_j)}{J} \tag{11}$$

$$RMSE(\hat{\lambda}) = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\lambda}_j - \lambda_j)^2}{J}} \tag{12}$$

where λ refers to an item parameter and J is the number of parameters in a parmeter type. Average bias and RMSE are the averages of these values across the number of replications. That is, average or mean bias and RMSE were obtained as the average difference between the estimated and "true" parameters across items of parameter types and replications.

Lastly, model fit indices of the log-likelihood (LL), Akaike information criteria (AIC), and Bayesian information criteria (BIC) were obtained and compared with particular focus on the differences between the different structural models across the simulation conditions.

Chapter IV

Results

4.1 Model Convergence Rates

The primary goal of the research was to examine the effects of missing data, specifically PM data designs, on the item parameter recovery of DCMs. For this purpose, data were generated and analyzed under two different PM designs of the two-form and BIB design which were each compared to each other and the condition of complete data.

Model convergence information was collected for each replication of each simulation condition. All models converged across all conditions of PM data design, number of attributes, structural model imposed, and sample size. In other words, the convergence rate was 100% for all simulation conditions.

4.2 Item Parameter Recovery Results

The average Bias and RMSE of the recovery of each parameter type (i.e., intercept and main effect) derived from 25 replications of each simulation condition are organized in Table 4. 1, Table 4. 2, Table 4. 3, and Table 4. 4. As expected, the condition of no missing data resulted in the lowest average bias and RMSE values and, thus best parameter recovery for both intercept and main effect parameters across all simulation conditions when compared to both types of PM data designs. Nonetheless, it was possible to see that item parameters of DCMs were also reasonably well recovered in most conditions even with missing data, and extremely well recovered in some of those conditions. There also seemed to be differences between the two types of PM designs; although they both generally showed similar trends across the conditions. Although the

amount of missingness was set equal per respondent, the two designs differed regarding whether they included common items as well as the number of forms. Among the two, results showed that using the BIB design resulted in a somewhat better recovery of overall model parameters when compared to the common item or two-form design. Furthermore, the degree to which the parameter recovery results of PM data designs agreed with the non-missing data condition differed depending on the other manipulated factors, which will be explored in more detail below. One noticeable thing regarding the item parameter recovery of the two types of parameters was that bias was always positive (or near zero) for intercept parameters while always negative (or near zero) for main effect parameters. RMSE of the main effect parameters were always larger than their intercept counterparts. More detail can be found in the upcoming sections.

Table 4. 1 Average Bias of Intercept Parameters

Sample Size	Model	Attribute	Two-Form	BIB	None
600	Saturated	4	0.103	0.076	0.066
600	Saturated	8	0.207	0.174	0.038
600	Log-Linear	4	0.080	0.044	0.056
600	Log-Linear	8	0.013	0.020	0.008
600	Higher-Order	4	0.078	0.038	0.055
600	Higher-Order	8	0.009	0.022	0.002
1200	Saturated	4	0.102	0.077	0.064
1200	Saturated	8	0.065	0.086	0.035
1200	Log-Linear	4	0.077	0.061	0.055
1200	Log-Linear	8	0.032	0.037	0.019
1200	Higher-Order	4	0.077	0.066	0.057
1200	Higher-Order	8	0.026	0.036	0.018

Sample Size	Model	Attribute	Two-Form	BIB	None
4800	Saturated	4	0.042	0.034	0.043
4800	Saturated	8	0.015	0.018	0.002
4800	Log-Linear	4	0.029	0.025	0.035
4800	Log-Linear	8	0.010	0.010	0.003
4800	Higher-Order	4	0.031	0.027	0.035
4800	Higher-Order	8	0.008	0.011	0.003

Table 4. 2 Average RMSE of Intercept Parameters

Sample Size	Model	Attribute	Two-Form	BIB	None
600	Saturated	4	0.354	0.374	0.249
600	Saturated	8	0.684	0.563	0.214
600	Log-Linear	4	0.337	0.345	0.249
600	Log-Linear	8	0.296	0.310	0.190
600	Higher-Order	4	0.339	0.337	0.251
600	Higher-Order	8	0.290	0.310	0.185
1200	Saturated	4	0.275	0.272	0.215
1200	Saturated	8	0.222	0.229	0.137
1200	Log-Linear	4	0.262	0.265	0.210
1200	Log-Linear	8	0.193	0.190	0.128
1200	Higher-Order	4	0.266	0.263	0.211
1200	Higher-Order	8	0.186	0.187	0.128
4800	Saturated	4	0.201	0.195	0.185
4800	Saturated	8	0.101	0.103	0.069
4800	Log-Linear	4	0.197	0.192	0.183
4800	Log-Linear	8	0.098	0.097	0.068
4800	Higher-Order	4	0.200	0.194	0.184

Sample Size	Model	Attribute	Two-Form	BIB	None
4800	Higher-Order	8	0.097	0.097	0.067

Table 4. 3 Average Bias of Main Effect Parameters

Sample Size	Model	Attribute	Two-Form	BIB	None
600	Saturated	4	-0.182	-0.176	-0.057
600	Saturated	8	-0.475	-0.355	-0.089
600	Log-Linear	4	-0.170	-0.169	-0.061
600	Log-Linear	8	-0.077	-0.056	-0.033
600	Higher-Order	4	-0.179	-0.160	-0.059
600	Higher-Order	8	-0.050	-0.039	-0.017
1200	Saturated	4	-0.101	-0.081	-0.049
1200	Saturated	8	-0.131	-0.145	-0.063
1200	Log-Linear	4	-0.114	-0.084	-0.053
1200	Log-Linear	8	-0.059	-0.062	-0.039
1200	Higher-Order	4	-0.111	-0.083	-0.052
1200	Higher-Order	8	-0.044	-0.050	-0.029
4800	Saturated	4	-0.030	-0.040	-0.023
4800	Saturated	8	-0.034	-0.037	-0.010
4800	Log-Linear	4	-0.035	-0.044	-0.026
4800	Log-Linear	8	-0.024	-0.026	-0.011
4800	Higher-Order	4	-0.036	-0.042	-0.025
4800	Higher-Order	8	-0.013	-0.018	-0.005

Table 4. 4 Average RMSE of Main Effect Parameters

Sample Size	Model	Attribute	Two-Form	BIB	None
4	Saturated	600	0.741	0.825	0.517
8	Saturated	600	1.332	0.999	0.339
4	Log-Linear	600	0.715	0.810	0.524
8	Log-Linear	600	0.474	0.476	0.292
4	Higher-Order	600	0.724	0.791	0.522
8	Higher-Order	600	0.422	0.439	0.276
4	Saturated	1200	0.594	0.617	0.532
8	Saturated	1200	0.364	0.341	0.209
4	Log-Linear	1200	0.610	0.617	0.532
8	Log-Linear	1200	0.292	0.267	0.193
4	Higher-Order	1200	0.602	0.617	0.531
8	Higher-Order	1200	0.277	0.255	0.186
4	Saturated	4800	0.498	0.514	0.486
8	Saturated	4800	0.155	0.143	0.098
4	Log-Linear	4800	0.499	0.514	0.486
8	Log-Linear	4800	0.149	0.138	0.099
4	Higher-Order	4800	0.741	0.825	0.517
8	Higher-Order	4800	1.332	0.999	0.339

4.2.1 Results by Number of Attributes

Along with the PM designs's impact on the item parameter recovery of DCMs, their relationships with the other manipulated factors of the number of attributes, structural model, and sample size were of interest. Graphical representations of the average bias and RMSE values with a focus on the number of attributes are provided in Figure 4. 1. The results show that for most conditions, the cases with eight attributes showed better parameter recovery. The exception was the eight attribute cases where the data were estimated using a saturated model, which showed the worst performance in parameter recovery across all conditions when data was missing (for both BIB and two-form designs). The non-missing data conditions displayed better parameter recovery compared to the missing data cases and more so for the four attribute conditions as opposed to the eight attribute conditions. Comparing within missingness conditions, we find that there is less difference between the two PM designs relative to the complete data case. While there is some preference for the BIB design over the two-form design in the four attribute conditions, this preference is not so in the eight attribute conditions. In fact, the reverse happens in some eight attribute conditions where the two-form design has lower bias and RMSE values compared to the BIB design.

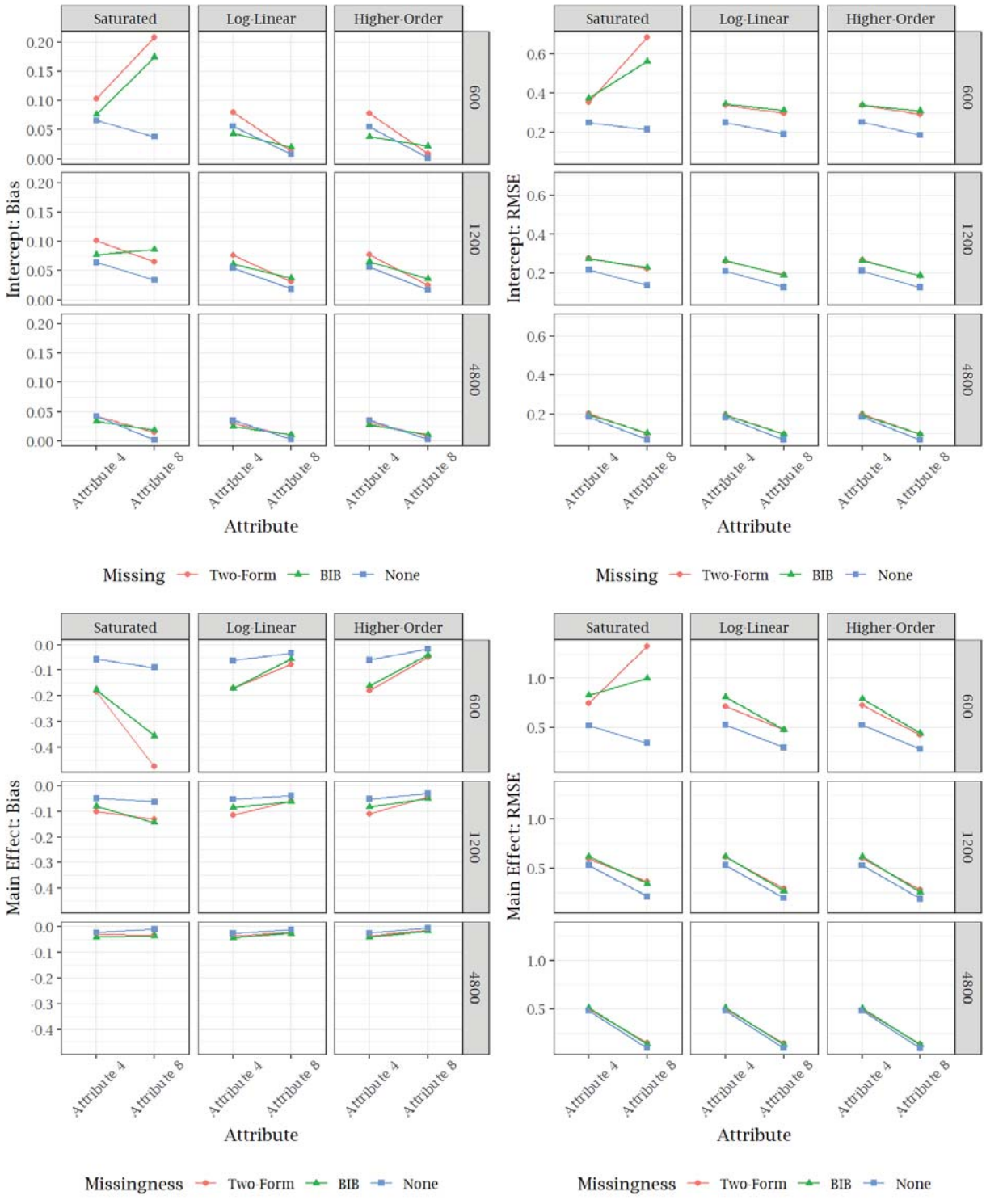

Figure 4. 1 Item Parameter Recovery Results of PM Data Designs focusing on the Number of Attributes

4.2.2 Results by Type of Structural Model

Plots of the average bias and RMSE values for item parameters with a focus on the structural part of DCMs are depicted in Figure 4. 2. Overall, the results indicate that imposing a more structured model on the structural part, be it as a log-linear model or high-order model, results in better item parameter recovery across conditions. Within each type of structural model, although the complete data case displayed better parameter recovery over the PM designs in almost all conditions, the results showed that excluding conditions for the saturated structural model for the eight attribute case with small sample size, model parameters were adequately recovered across the conditions even with missing data, with performance improving as sample size was increased. Most noticeable are the particularly pronounced bias and RMSE values for item parameters for the eight attribute case when a saturated model is used in conjunction with small sample sizes in the case of missing data. That is, it seemed that the choice of the structural model for DCMs had a greater impact when data were missing with increasing attribute sizes and decreasing sample sizes. Although the complete data also showed better recovery when a structure was imposed, there was much less difference in the results. There was also some preference for the BIB design over the two-form design for each structural model of DCMs. Furthermore, it was possible to see that the two-form design was more impacted by the structural model than the BIB design. Elaborating, the former design showed better performance in terms of parameter recovery when a higher-order model was imposed relative to a log-linear model. The BIB design did not seem to prefer either structural model.

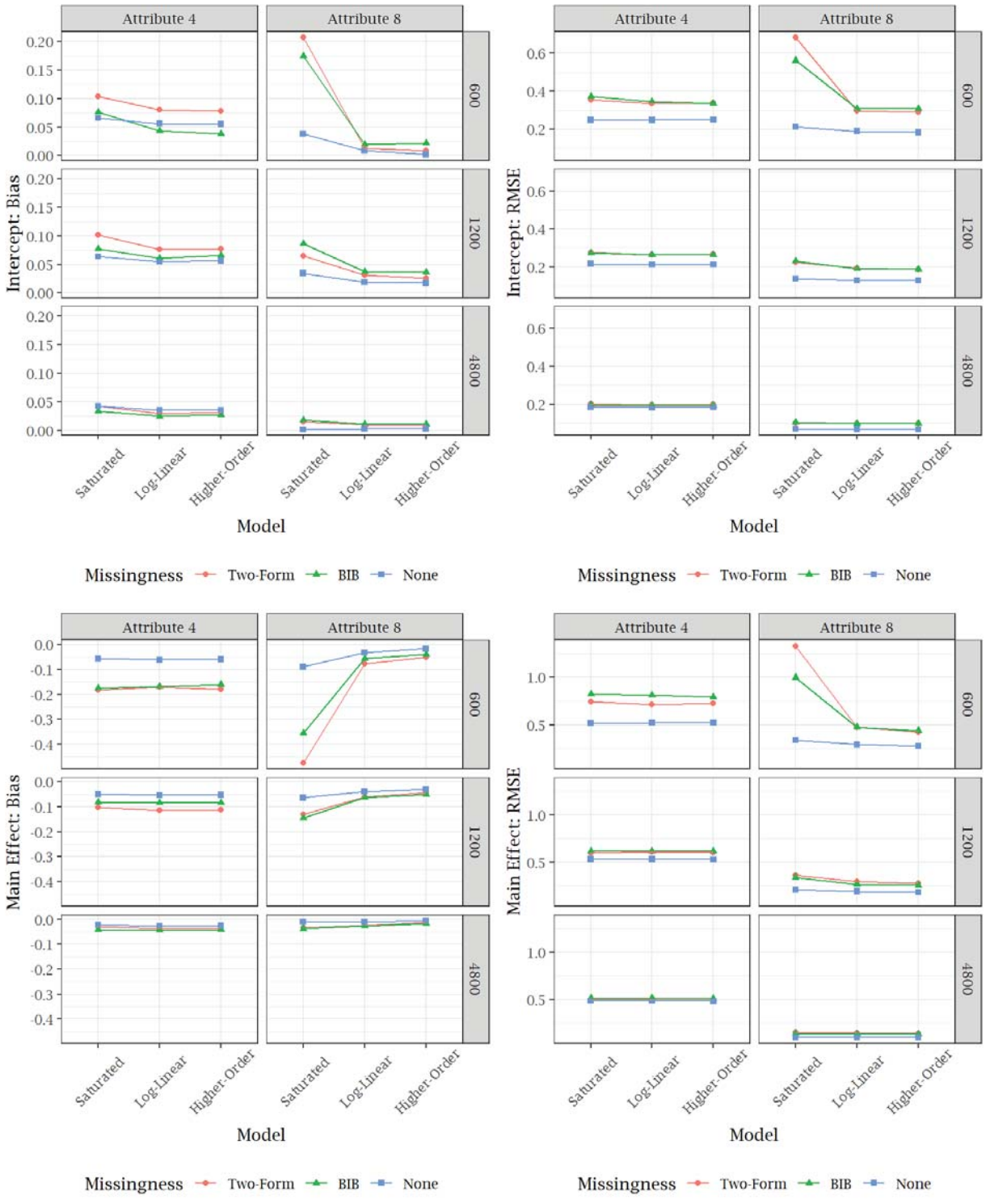

Figure 4. 2: Item Parameter Recovery Results of PM Data Designs focusing on the Number of Type of DCM Structural Model

4.2.3 Results by Sample Size

The results of average bias and RMSE for parameter recovery with focus on the sample size are graphically presented in Figure 4. 3. It is evident that parameter recovery improves as sample size increases. This trend is regardless of the type of missingness design (that is, whether there was missing data or not). However, the results also show that increasing sample sizes impacts item parameter recovery more for cases with missing data relative to cases where no missing data were assumed. The most obvious instance is the missing data cases where the saturated model was the structural model with eight attributes. There was a large gap between going from sample size of 600 and 1200, which did not appear in the complete data cases. Furthermore, there was more difference in parameter recovery going from small to average sample sizes than moving from average to large sample sizes for the same cases. Also, the results implied that the BIB design performed better than the two-form design for these scenarios. The results also found that BIB design had better recovery than the two-form design for four attributes and sample sizes of 600 and, to a lesser extent, 1200. While we could see interaction effects between sample size and number of attributes as well as the saturated structural model, there didn't seem to be much difference between the log-linear and higher-order structural models depending on sample size. In the case of very large sample sizes, the parameter recovery across all missingness conditions was extremely comparable, even in the case a saturated structural model was fit.

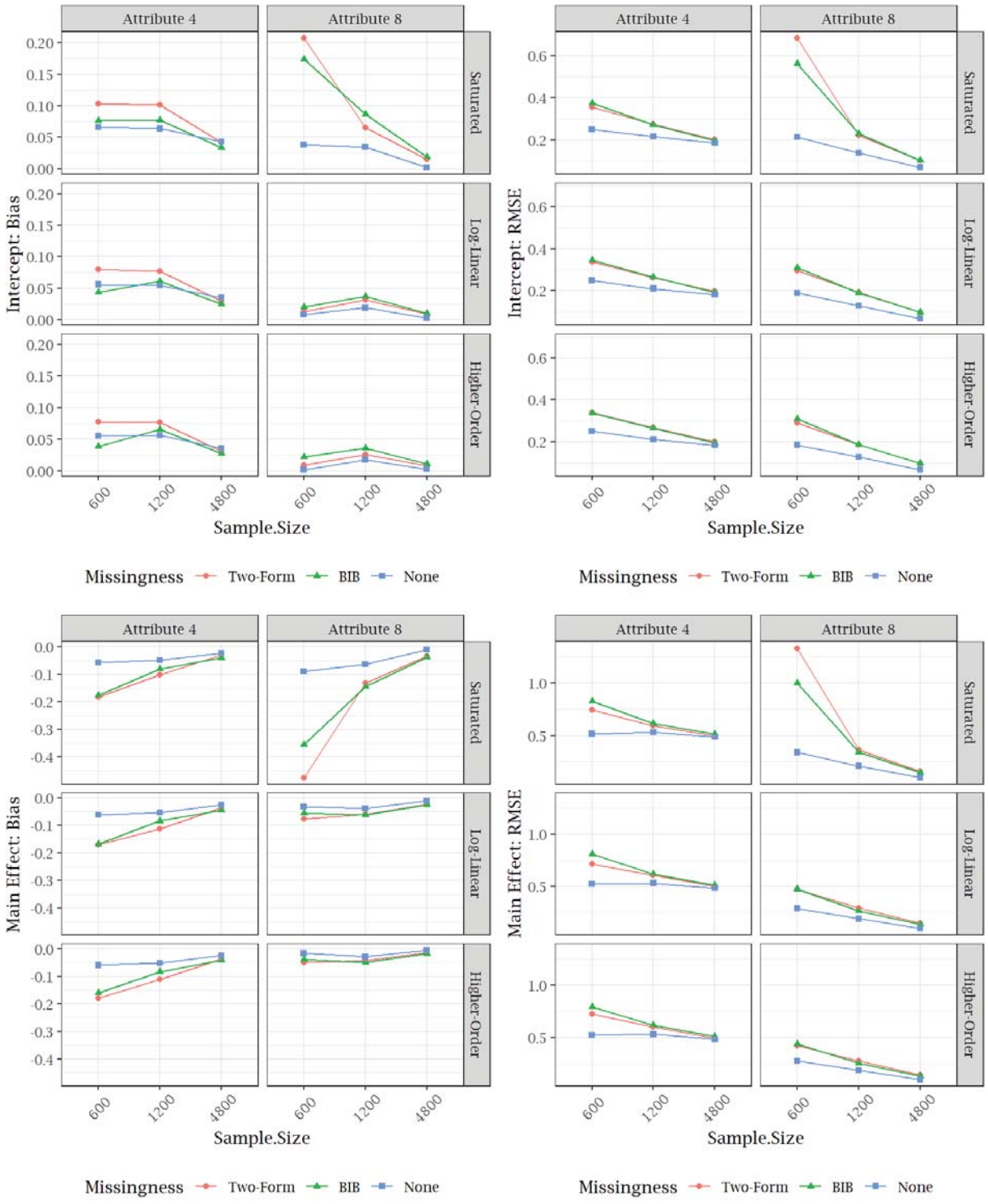

Figure 4. 3: Item Parameter Recovery Results of PM Data Designs focusing on the Number of Type of DCM Structural Model

4.3 Model Fit Results

Model fit across simulation conditions was assessed using the log-likelihoods, AIC and BIC indices. The focus was on comparing the different structural models across conditions. The average of model fit results across replications are organized in Table 5. 5, Table 5.6, and Table. 5.7 for all the simulation conditions. Both the log-linear structural model and the higher-order structural model is nested within the saturated structural model so it is possible to conduct likelihood ratio tests between the satuarated and either of the log-linear and higher-order models. The goal was to see whether the reduction in model parameters do to imposing constraints on the structural model significantly impacted model fit. Likelihood ratio test results across all replications of all simulation conditions did not show a single case where the saturated model was preferred over the log-linear or higher-order model. That is, imposing a log-linear or higher-order model did not significantly impact model fit.

AIC and BIC indices corroborated such results where the AIC and BIC values for the log-linear and higher-order structural models were always smaller compared to those for the saturated model. AIC and BIC was also used to compare the relative model fit between the log-linear model and the higher-order model. In all cases, the higher-order structural model had lower AIC and BIC values compared to the log-linear model. In short, the AIC and BIC values could be ordered from saturated, log-linear, and higher-order structural model. There were large differences in the AIC and BIC values for the staturated and either the log-linear and higher-order model. The differences between the loge-linear and higher-order model was much smaller but still the higher-order model had the smaller AIC and BIC values. That is, according to these relative fit indices, the higher-order model was most preferred. Considering that the log-linear model consisted

of sixe parameters, thre first-order and three second-order parameters, and the higher-order model was a 1PL model with only four parameters (one slope parameter and three intercept parameters), this implies that the higher-order model does the same job using a lesser number of parameters. Thus, the higher-order structural model may be the preferable of the two.

Table 4. 5 Average Log-Likelihood Values across Simulation Conditions

Attribute	Sample.Size	Model	Two-Form	BIB	None
4	600	Saturated	3,035.971	3,058.562	4,445.208
4	600	Log-Linear	3,038.671	3,061.367	4,448.253
4	600	Higher-Order	3,041.707	3,063.194	4,451.475
4	1200	Saturated	6,090.095	6,135.161	8,916.844
4	1200	Log-Linear	6,093.618	6,138.172	8,921.616
4	1200	Higher-Order	6,097.375	6,141.297	8,925.917
4	4800	Saturated	24,462.166	24,611.342	35,769.049
4	4800	Log-Linear	24,468.616	24,616.785	35,776.888
4	4800	Higher-Order	24,474.766	24,623.523	35,785.422
8	600	Saturated	5,943.649	5,966.721	8,712.349
8	600	Log-Linear	6,000.263	6,022.441	8,785.182
8	600	Higher-Order	6,013.797	6,036.116	8,796.199
8	1200	Saturated	11,923.875	11,990.468	17,469.844
8	1200	Log-Linear	11,991.348	12,054.444	17,554.996
8	1200	Higher-Order	12,005.986	12,066.232	17,566.313
8	4800	Saturated	48,116.396	48,289.433	70,365.824
8	4800	Log-Linear	48,208.789	48,381.227	70,481.651
8	4800	Higher-Order	48,215.517	48,386.990	70,479.023

Table 4. 6 Average AIC Values across Simulation Conditions

Attribute	Sample.Size	Model	Two-Form	BIB	None
4	600	Saturated	6,149.942	6,195.124	8,968.416
4	600	Log-Linear	6,145.341	6,190.734	8,964.505
4	600	Higher-Order	6,141.415	6,184.388	8,960.951
4	1200	Saturated	12,258.190	12,348.321	17,911.688
4	1200	Log-Linear	12,255.237	12,344.343	17,911.232
4	1200	Higher-Order	12,252.750	12,340.595	17,909.834
4	4800	Saturated	49,002.331	49,300.685	71,616.098
4	4800	Log-Linear	49,005.232	49,301.570	71,621.777
4	4800	Higher-Order	49,007.533	49,305.046	71,628.844
8	600	Saturated	12,493.299	12,539.443	18,030.697
8	600	Log-Linear	12,168.526	12,212.883	17,738.364
8	600	Higher-Order	12,141.594	12,186.232	17,706.398
8	1200	Saturated	24,453.750	24,586.935	35,545.687
8	1200	Log-Linear	24,150.695	24,276.888	35,277.993
8	1200	Higher-Order	24,125.972	24,246.464	35,246.625
8	4800	Saturated	96,838.791	97,184.867	141,337.648
8	4800	Log-Linear	96,585.578	96,930.453	141,131.302
8	4800	Higher-Order	96,545.033	96,887.979	141,072.046

Table 4. 7 Average BIC Values across Simulation Conditions

Attribute	Sample.Size	Model	Two-Form	BIB	None
4	600	Saturated	6,321.422	6,366.604	9,139.896
4	600	Log-Linear	6,294.837	6,340.230	9,114.001
4	600	Higher-Order	6,268.926	6,311.899	9,088.462
4	1200	Saturated	12,456.703	12,546.834	18,110.201
4	1200	Log-Linear	12,428.299	12,517.406	18,084.295
4	1200	Higher-Order	12,400.362	12,488.207	18,057.446
4	4800	Saturated	49,254.910	49,553.263	71,868.677
4	4800	Log-Linear	49,225.428	49,521.767	71,841.973
4	4800	Higher-Order	49,195.347	49,492.860	71,816.658
8	600	Saturated	13,825.569	13,871.712	19,362.967
8	600	Log-Linear	12,537.868	12,582.225	18,107.706
8	600	Higher-Order	12,392.219	12,436.857	17,957.023
8	1200	Saturated	25,996.043	26,129.229	37,087.980
8	1200	Log-Linear	24,578.262	24,704.455	35,705.559
8	1200	Higher-Order	24,416.107	24,536.598	35,536.760
8	4800	Saturated	98,801.132	99,147.207	143,299.989
8	4800	Log-Linear	97,129.593	97,474.468	141,675.318
8	4800	Higher-Order	96,914.187	97,257.132	141,441.199

Chapter V

Discussion

5.1 Summary of Results

Recently, DCMs and their capability to classify respondents into their most probable mastery profiles and provide diagnostic feedback have been gaining ground, particularly in education. While research on DCMs has been very active, there is a dearth of studies on the impact of missing data on DCMs. Missing data is almost inevitable in assessments and studies, and it can cause various problems for models and their consequent conclusion if not appropriately accounted for. While missing data is viewed as generally unfavorable, strategically incorporating missing data in the form of PM designs can have many benefits, which is boosted due to the improvement of techniques of handling missing data that can be assumed to be at least MAR. In the field of education, especially in large-scale assessments, PM designs are appealing as they help alleviate concerns about data collection costs, respondent fatigue, and data quality by allowing many items and variables to be collected while reducing the number of individuals having to respond to each item (Gonzalez & Rutkowski, 2010).

PM designs can be helpful in DCMs applications which are increasingly coming to involve more and more attributes. For model identification and stable estimation of model parameters, this means that the number of items per attribute must increase as well. In addition, the likelihood of DCMs being used in large-scale settings is increasing following the need for diagnostic feedback. Accordingly, this thesis aimed to explore the effects of missing data with a focus on two PM designs based on matrix sampling on the parameter recovery of DCMs. The two PM designs considered were the two-form design

and the BIB design. The impact of various factors in DCM estimation, specifically the number of attributes, structural model formulation, and sample size was examined in relation to the PM design. The results were compared in terms of the recovery of "true" parameters and to corresponding results from DCMs with no missing data.

Results showed that while the complete data cases resulted in the best parameter recovery across the simulation conditions, the item parameter recovery in the cases of missing data under both PM designs also ranged from adequate to extreme good for most conditions. Elaborating, parameter recovery improved as the number of attributes increased, when constraints were imposed on the structural part of DCMs, and as sample size increased. Out of the three factors, the most salient was the effect of sample size, followed by choice of the structural model, and lastly, attributes.

Large sample sizes were found to be the most important, resulting in the best parameter recovery not only in terms of each missingness design (i.e., none, two-form design, and BIB design) but also resulting in nearly identical results across missingness designs. It could even compensate for using a saturated structural model. Such large sample sizes like 4800 might not be feasible in real-life settings, and the results show that an average sample size of around 1000 is also favorable. Nonetheless, researchers should aim to recruit as many participants as possible. Furthermore, when the sample size is small, they must be more careful with other DCM design factors, such as the structural model used.

In terms of the structural model factors, reducing the number of structural parameters to be estimated by setting a certain model had lower bias and RMSE values overall, particularly in cases of small sample sizes. There wasn't much difference in the choice of a log-linear structural model relative to a higher-order one, however, the two-

form design slightly preferred the latter. However, the log-linear model in this study consisted of six parameters of the first and second-order parameters, while the higher-order model was a 1PL model with only four parameters. This implied that the higher-order model does the same job using a lesser number of parameters. Thus, the higher-order structural model may be the preferable of the two. Relative model fit indices of the AIC and BIC corroborated such results.

It was somewhat surprising that increasing the number of attributes led to better parameter recovery. This is perhaps more so the effect of the increase in the number of items following the increase of attributes. Nevertheless, the number of attribute factors showed the largest variability in conjunction with other factors to show us the importance of having a clear structural model when the number of attributes is large and sample sizes are small.

Regarding the two PM data designs, they both mostly displayed similar trends across the simulation conditions. However, the BIB design was preferred in more conditions than the two-form design. This difference in parameter recovery, even though the amount of missingness was set equal per respondent, points to the fact that although the amount of missing data itself is important, the specific PM design used, whether it be in terms of the number of forms used or the types of items allocated, also matters. Thus, the effects of the PM design need to also be carefully investigated, and the best design chosen.

BIBLIOGRAPHY

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale assessments in education*, *1*(1), 1-27.

Betini, G. S., Avgar, T., & Fryxell, J. M. (2017). Why are we not evaluating multiple competing hypotheses in ecology and evolution?. *Royal Society Open Science*, *4*(1), 160756.

Bradshaw, L. (2017). Diagnostic classification models. *The handbook of cognition and assessment: Frameworks, methodologies, and applications*, 297-327.

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99-118.

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, *3*, 297-321.

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. In *Frontiers in Education* (p. 372). Frontiers.

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199.

De la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353.

Duck-Mayr, J., Garnett, R., & Montgomery, J. (2020, August). GPIRT: A Gaussian Process Model for Item Response Theory. In *Conference on Uncertainty in Artificial Intelligence* (pp. 520-529). PMLR.

Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika*, *84*(1), 19-40.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*(1), 128-141.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, *3*, 125-156.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological methods*, *11*(4), 323.

Hansen, M. P. (2013). *Hierarchical item response models for cognitive diagnosis*. (Doctoral dissertation, UCLA).

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191.

Hogue, C. M., Pornprasertmanit, S., Fry, M. D., Rhemtulla, M., & Little, T. D. (2013). Planned missing data designs for spline growth models in salivary cortisol research. *Measurement in Physical Education and Exercise Science*, *17*(4), 310-325.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1-17.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, *26*(1), 031-73.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*(5), 850-874.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59-81.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Little, T. D., Gorrall, B. K., Panko, P., & Curtis, J. D. (2017). Modern practices to improve human development research. *Research in Human Development*, *14*(4), 338-349.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of pediatric psychology*, *39*(2), 151-162.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*(4), 199-204.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247-264.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *163*(3), 445-459.

Nakagawa, S. (2017). Missing data: Mechanisms, methods, and messages. In Fox G. A., Negrete-Yankelevich S., & Sosa V. J. (Eds.), *Ecological statistics: Contemporary theory and application* (pp. 81–105). Oxford University Press.

Noble, D. W., & Nakagawa, S. (2021). Planned missing data designs and methods: Options for strengthening inference, increasing research efficiency and improving animal welfare in ecological and evolutionary research. *Evolutionary Applications*, *14*(8), 1958-1968.

O'muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(2), 177-194.

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24-56.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in irt models. *Psychometrika*, *82*(3), 795–819. https://doi.org/10.1007/s11336-016-9544-7

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), i-53.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219-262.

Rupp, AA., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Ballinger.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of educational measurement*, 345-354.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317-339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*(3), 287.

Thompson, W. (2018). *Evaluating model estimation processes for diagnostic classification models* (Doctoral dissertation, University of Kansas).